Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 11 (2012) 55 – 62

Procedia Computer Science

Proceedings of the 3rd International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2012)

Bibliome mining platform and application for building metabolic interaction network

P. Patumcharoenpol[a,b], J. Chan[b,c], A. Meechai[b,d,e], B. Shen[a], W. Vongsangnak[a]*

[a]*Center for Systems Biology, Soochow University, No. 1. Shizi Street, Suzhou 215006, China*
[b]*Bioinformatics and Systems Biology Program,* [c]*School of Information Technology,*
[d]*Department of Chemical Engineering, and* [e]*Biological Engineering Program*
King Mongkut's University of Technology Thonburi, Bangkok, Thailand

Abstract

Bibliome mining is a powerful technique for large-scale information extraction from textual data and connecting between biological entities as well as functional hypotheses. Currently, most bibliome mining is used for some specific studies involving genes and proteins; however, much less efforts have been focused on metabolites. In addition to application works, the focus has been on proving the concepts and algorithms, but very few reports on development of applicable text mining platform. In this study, we aimed to develop a bibliome mining platform that could be used to perform basic text mining tasks and further be used for building metabolic interaction networks. We developed a platform with the evaluated tools and subsequently tested its functions for extraction of interactions between biomolecules (e.g. genes, enzymes, proteins and metabolites) in yeast *Saccharomyces cerevisiae*. The results were then manually curated afterwards using KEGG LIGAND and public yeast database. Of the collected 11 text mining tools, we selected 3 suitable tools, namely ABNER, OpenNLP and LingPipe for further implementing the bibliome mining platform. In summary, a prototype of a bibliome mining platform was successfully developed and may further be used for building metabolic interaction network of *S. cerevisiae*. This study can be used as a basic framework for further improvement as well as extension of relevant text mining tasks and broad applications.

Selection and/or peer-review under responsibility of the Program Committee of CSBio 2012.

Keywords: bibliome mining, biomolecules, metabolic interaction network, *Saccharomyces cerevisiae*, text mining

* Corresponding author. Tel.: +86 18051095601.
E-mail address: wanwipa@suda.edu.cn .

doi:10.1016/j.procs.2012.09.007

1. Introduction

Up to date, there are various kinds of biological knowledge which were successfully derived from the texts, such as protein-protein interaction [1], metabolite-enzyme interaction [2], regulated pathway [3], and also gene-disease relationship [4]. Undoubtedly, bibliome mining is thus powerful technique. Recently, there have been studies on development of bibliome mining pipeline assembler with graphical user interface (GUI) for integration of several text mining tools, such as Argo, U-compare, and @Note [5–7] for biomedical texts. Even though these tools provided GUI for user with ease of use, they were tightly coupled with its existing pipeline assembler and this consequently resulted in difficulty of tools expansion, and modification, which leads to limited application. For examples, text mining research community focused on genes, proteins, and metabolites annotation [8,9]. In addition, most of the studies only focused on proving the concepts and algorithms in text mining, but much less studies on developing text mining platform for a general application [7], such as building interaction between genes, enzymes, proteins, metabolites as biomolecules representing in the form of metabolic network. In general, building of metabolic network of different organisms (e.g. bacteria, yeast, fungi and human) was based on genome databases and bioinformatics tools for functional annotation [10,11]. On the other hand, if no genome information available, it was mainly rely on literature for metabolic network reconstruction [12]. Certainly, manual curation is performed for identification of gene, enzyme, protein and metabolite relationship and further construction of the metabolic network. Once concerning to automatic approach and time consuming, bibliome mining is therefore considered to be an alternative approach for building metabolic interaction network.

In this study, we aim to develop bibliome mining platform that could be used as an initial framework to perform basic text mining tasks and further be used for building metabolic interaction network. To explore, we initially collected text mining tools that were publicly available and further evaluated based on their features and F-scores. Once evaluation process was finished, we then selected suitable text mining tools for development of bibliome mining platform. Through the end, we tested the developed platform using yeast *Saccharomyces cerevisiae* texts for building its metabolic interaction network and manually curated the network using metabolic pathway database of *S. cerevisiae* (KEGG LIGAND) and public yeast database.

2. Methods

An overall workflow for developing bibliome mining platform was divided into three main steps as illustrated in Fig. 1 below: 1) Collection, classification, selection and evaluation of text mining tools, 2) Development of bibliome mining platform, 3) Testing, curation and using the developed platform.

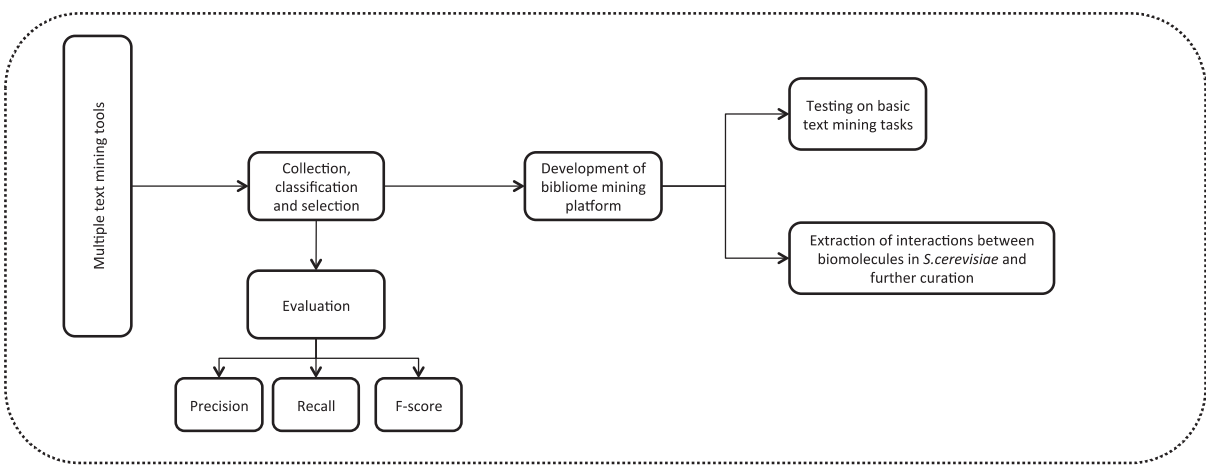

Fig. 1. Overall methodology of developing bibliome mining platform.

2.1. Collection, classification, selection and evaluation of text mining tools

Text mining tools were collected from different public resources. Each tool was reviewed and then classified under different categories: algorithms, licenses, Application Program Interface (API), custom model usage (machine learning algorithm), and capabilities. According to these classified categories, we then selected the tools based main categories, which had entity recognition capability and custom model usage. In addition, the selected tools were then evaluated using F-score with respect to detection of gene name, enzyme name, protein name, and metabolite name appeared in the texts. For these tasks, the tools were evaluated using 10 fold-cross validations on corpus from literature [13] and the corpus from BioCreative II GM [14]. In each corpus, recall and precision were firstly calculated and then F-score was then calculated. The formula of precision (1), recall (2) and F-score (3) are presented as shown in the following.

$$Precision = True\ positive/(True\ positive + False\ positive) \quad (1)$$

$$Recall = True\ positive/(True\ positive + False\ negative) \quad (2)$$

$$F\text{-}score = (2 \times Precision \times Recall)/(Precision + Recall) \quad (3)$$

2.2. Development of bibliome mining platform

We used Unstructured Information Management Architecture (UIMA) [15], which is an open source framework for analysis of an unstructured data, for main platform's implementation. For a design of platform, we separated analysis's step into three levels, namely Reader, Processor, and Consumer. Reader could transform texts into Common Analysis Structure (CAS). Processor could read and modify CAS and further allow to be operated with other processors. Consumer was able to read the CAS in order to report the results. In addition, we implemented processor module for detecting the interaction between biomolecules (e.g. genes, proteins, enzymes, metabolites) by co-occurrences in the texts. With any possibility of two or more biomolecules were mentioned within the same sentences, they were counted as interacting pairs. This module afterwards aggregated all potential input texts in order to obtain all possible interacting pairs as well as reported the frequency of co-occurrences of any interacting pairs.

2.3. Testing, curation and using the developed platform

The developed platform was tested for extraction of interactions between gene, enzyme, protein and metabolite of *S. cerevisiae*, which were taken from MEDLINE as a repository of biomedical text citation. The citation includes author, date, title and abstract. To screen the citation involved in *S. cerevisiae*, Medical Subject Heading (MESH) term was used for filtering and the abstract/title was then extracted. After that, we chopped the whole text in each citation to be an individual sentence. Subsequently, each sentence was annotated by entity recognition module, which was trained by CNobata2011 for metabolite and BioCreative II GM data for gene, enzyme and protein. Later, the relationship extraction module was used to collect all possible names of gene, enzyme, protein and metabolite in sentence and then annotated them as an interaction pair. Finally, all interaction pairs were aggregated from each sentence and sorted by their co-occurrences frequency. We considered the top 100 interaction pairs with more than 65 co-occurrences. The top 100 interaction pairs were then compared to a metabolic graph from KEGG LIGAND using graph walking. If any interaction pair obtained from MEDLINE matched with any interaction pair from KEGG LIGAND, then the results were marked as metabolite-enzyme interaction, metabolite-metabolite interaction, or enzyme-enzyme

interaction. In case of gene-gene interaction, gene-protein interaction, or protein-protein interaction, we were manually inspected using other public yeast databases (e.g. PUBMED and *Saccharomyces* Genome Database).

3. Results and Discussion

3.1. Features and F-scores assessment of selected text mining tools

At the beginning, we collected 11 text mining tools as listed in Table 1. Afterwards, we selected the tools based main features categories, which had entity recognition capability and could be able to perform custom model usage (machine learning algorithm). Under these considerations, 3 suitable text mining tools were selected, namely ABNER, OpenNLP, and LingPipe.

Once evaluation processes were performed, we assessed F-score of selected text mining tools with respect to detection of gene name, enzyme name, protein name, and metabolite name appeared in the texts. The corpus from literature CNobata2011 [13] containing metabolite annotation and the corpus from BioCreative II GM containing gene, enzyme, protein annotation were used. Interestingly, the results showed that ABNER obtained the highest F-score (66%) while OpenNLP obtained the lowest F-score (63%) for metabolite name detection. Concerning to gene, enzyme and protein detection, the results clearly showed that ABNER had the highest F-score (78%) whereas OpenNLP had the lowest F-score score (38%). These could give some clue that OpenNLP seems not suitable for detecting gene, enzyme and protein names. The results are shown in Table 2.

3.2. Implementation of bibliome mining platform

The list of modules was implemented as illustrated in Fig. 2. We implemented parser for BioCreative II, MEDLINE parser, and Folder Reader as a Reader. BioCreative II parser and MEDLINE parser used for parsing data from BioCreative II GM and MEDLINE, respectively. Folder reader read the data as a plain text from the folder without any processing. ABNER, LingPipe, and OpenNLP, and relationship extraction module were the processors for analysing text. We implemented some NLP modules from OpenNLP and LingPipe, separately (i.e. sentence detection, tokenization, etc.). Aggregation engine, Diagnostic Engine and Relationship Aggregation were implemented as the Consumers. Aggregation Engine and Diagnostic Engine were the reporters, which were used in statistical report for internal testing. Aggregation Engine combined the results from various documents. Diagnostic Engine compared the results within document.

3.3. Testing and using of bibliome mining platform

We built metabolic interaction network of yeast *S. cerevisiae* using our developed bibliome mining platform. We extracted 65,535 unique interaction pairs from 75,367 citations and manually inspected top 100 interaction pairs of biomolecules. Hereby we found 3 protein-protein interactions, 18 biomolecules complexes, 1 enzyme-metabolite interaction, 12 metabolite-metabolite interactions, 49 indirect interactions, and 17 false positive interactions. We illustrated interaction graph from interaction pairs among the list of top 100 interaction pairs with more than 65 co-occurrences by representing interaction network's modules with at least 3 nodes. As shown in Fig. 3, this graph shows logical connection between the nodes in metabolic context, such as carbon sources-based connection (e.g. glucose and fructose) and nitrogen sources-based connection (e.g. aspartate and glutamate). Besides, this graph also shows proteins complexes, e.g., SIR2/3/4 and MRE11/RAD50/XRS2. For other results identified with no current literature support, this was regarded as false positive interactions. For example, we often found false interactions between IgA and IgG in human immunity studies [16,17].

Table 1. Features comparison of 11 collected text mining tools.

Names	Pre-processing			Entity recognition		Post-processing		Others		
	Lemmatization	Shallow parse	Deep parse	Tagging	Customization	Relationship Extraction	API	License	Algorithm	Reference
ABNER				•	•		•	CPL	CRF**	[18]
OpenNLP		•		•	•		•	LGPL	CRF**	[19]
LingPipe		•		•	•		•	Free*	HMM**	[20]
Stanford NER				•			•	Free*	MEMM**	[21]
BioTagger				•			•	NA	CRF	[22]
Enju			•					NA	CFG-Filtering	[23]
GENIA Dependency parser			•					Free*	Modified LR-Algorithm	[24]
MorphaA	•							Free*	Finite state machine	[25]
BioLG		•					•	Free*	Rules-base	[26]
OpenDMap				•		•	•	MPL 1.1	DMAP	[27]
GenieTagger				•				Free*	Bidirectional Inference**	[28]

*Free under academic use, **Machine learning algorithm

Abbreviations: CRF: Conditional Random Field, HMM: Hidden Markov Model, MEMM: Maximal Entropy Markov Model, DMAP: Direct Memory Access Parsing. MPL: Mozilla Public License, CPL: Common Public License, LGPL: Less General Public License, NA: Not Available.

Table 2. Evaluation of text mining tools based on precision, recall and F-score.

Name	CNobata2011			BioCreative II GM		
	Precision	Recall	F-score	Precision	Recall	F-score
ABNER	77 %	58 %	66 %	84 %	74 %	78 %
LingPipe	61 %	67 %	64 %	60 %	79 %	68 %
OpenNLP	89 %	48 %	63 %	55 %	29 %	38 %

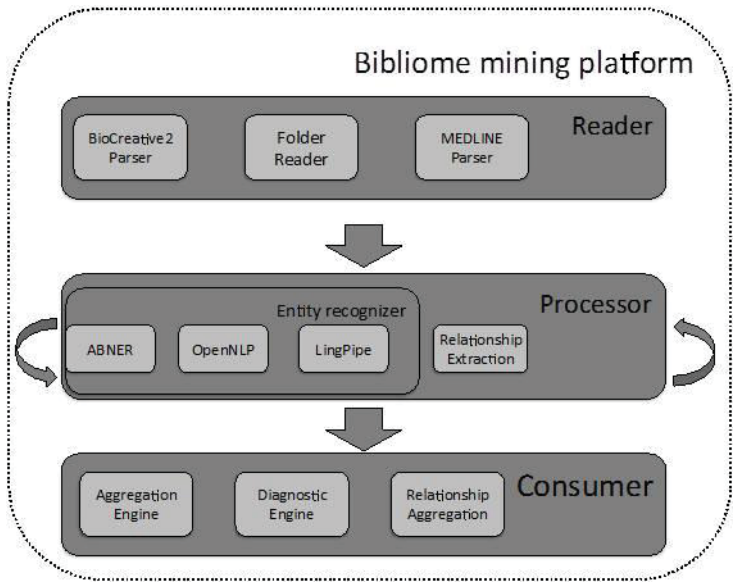

Fig. 2. Implementation of our bibliome mining platform.

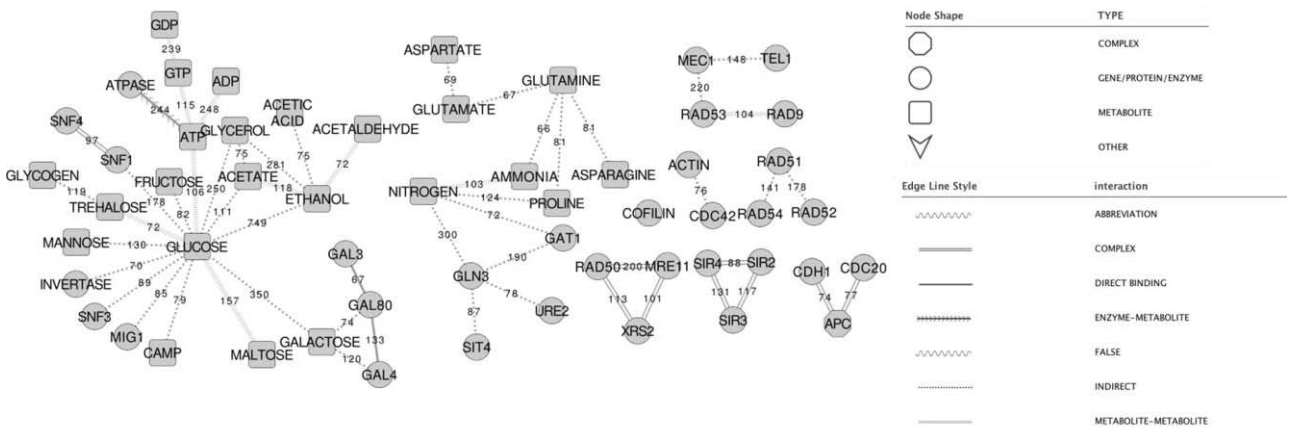

Fig. 3. The interaction graph from the top 100 interaction pairs with more than 65 co-occurrences annotated in *S. cerevisiae*'s texts. This graph shows interaction network's modules with at least 3 nodes.

4. Conclusions and Future work

The first prototype of bibliome mining platform was successfully developed and be used for performing basic text mining tasks (e.g. sentence detection and entity recognition). In addition, as shown in Fig. 3, our platform could be used for building metabolic interaction network of yeast *S. cerevisiae*.

Nonetheless, concerning to our developed platform, it remains some weaknesses. For examples, the platform is not easily used by biologists since it requires programming and technical skills for operation. Moreover, in terms of building of metabolic interaction network, the relationships or interactions between biomolecules are not yet specified. For future work, we plan to improve the platform that can automatically predict the specific type of interaction extracting metabolic interaction network by adding a relationship extraction module. In addition, we also propose to implement an intuitive user interface and release this tool as a web service.

Acknowledgements

Preecha Patumcharoenpol would like to thank National Center for Genetic Engineering and Biotechnology (BIOTEC), King Mongkut's University of Technology Thonburi (KMUTT), and the Soochow University for financial support. Wanwipa Vongsangnak is supported by initial funding from Soochow University.

References

[1] R. Chowdhary, J. Zhang, J.S. Liu. Bayesian inference of protein-protein interactions from biological literature. *Bioinformatics (Oxford, England)* 2009;**25**:1536–42.

[2] L. Zhang, D. Berleant, J. Ding, T. Cao, E. Syrkin Wurtele. PathBinder--text empirics and automatic extraction of biomolecular interactions. *BMC Bioinformatics* 2009;**10**.

[3] J. Saric, L.J. Jensen, R. Ouzounova, I. Rojas, P. Bork. Extraction of regulatory gene/protein networks from Medline. *Bioinformatics (Oxford, England)* 2006;**22**:645–50.

[4] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, D.S. Wishart. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research* 2008;**36**:W399–405.

[5] R. Rak, A. Rowley, W. Black, S. Ananiadou. Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database (Oxford)* 2012;**2012**:bas010.

[6] Y. Kano, W.A. Baumgartner, L. McCrohon, S. Ananiadou, K.B. Cohen, L. Hunter, et al. U-Compare: share and compare text mining tools with UIMA. *Bioinformatics (Oxford, England)* 2009;25:1997–8.

[7] A. Lourenço, R. Carreira, S. Carneiro, P. Maia, D. Glez-Peña, F. Fdez-Riverola, et al. @Note: a workbench for biomedical text mining. *Journal of Biomedical Informatics* 2009;**42**:710–20.

[8] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, J. Tsujii, Overview of BioNLP Shared Task 2011, ACL HLT 2011. 2009 (2011) 1–6.

[9] C.N. Arighi, Z. Lu, M. Krallinger, K.B. Cohen, W. Wilbur, A. Valencia, et al. Overview of the BioCreative III Workshop. *BMC Bioinformatics* 2011;**12**.

[10] A.M. Feist, M.J. Herrgård, I. Thiele, J.L. Reed, B.Ø. Palsson. Reconstruction of biochemical networks in microorganisms. *Nature Reviews. Microbiology* 2009;**7**:129–43.

[11] A. Bordbar, B.O. Palsson. Using the reconstructed genome-scale human metabolic network to study physiology and pathology. *Journal of Internal Medicine* 2012;**271**:131–41.

[12] M.R. Andersen, M.L. Nielsen, J. Nielsen. Metabolic model integration of the bibliome, genome, metabolome and reactome of Aspergillus niger. *Molecular Systems Biology* 2008;**4**:178.

[13] C. Nobata, P.D. Dobson, S.A. Iqbal, P. Mendes, J. Tsujii, D.B. Kell, et al. Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics* 2011;**7**:94–101.

[14] L. Smith, L.K. Tanabe, R.J. nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, et al. Overview of BioCreative II gene mention recognition. *Genome Biology* 2008;**9**.

[15] D. Ferrucci, A. Lally. *Towards an interoperability standard for text and multi-modal analytics*. IBM Research Report; 2006.
[16] M. Albrechtsen, G.R. Yeaman, M.A. Kerr. Characterization of the IgA receptor from human polymorphonuclear leucocytes. *Immunology* 1988;64:201–5.
[17] C.L. Sutton, H. Yang, Z. Li, J.I. Rotter, S.R. Targan, J. Braun. Familial expression of anti-*Saccharomyces cerevisiae* mannan antibodies in affected and unaffected relatives of patients with Crohn's disease. *Gut* 2000;46:58–63.
[18] B. Settles. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* 2005;21:3191–2.
[19] G. Wilcock. Text Annotation with OpenNLP and UIMA. *Proceedings of the 17th Nordic Conference of Computational Linguistics* 2009;:7–8.
[20] B. Carpenter. LingPipe for 99.99 % Recall of Gene Mentions. *Proceedings of the Second BioCreative Challenge* 2007;:2–4.
[21] S. Batra, D. Rao, Entity Based Sentiment Analysis on Twitter, Science. (2010) 1–12.
[22] Y. Jin, R.T. McDonald, K. Lerman, M.A. Mandel, S. Carroll, M.Y. Liberman, et al. Automated recognition of malignancy mentions in biomedical literature. *BMC Bioinformatics* 2006;7:492.
[23] Y. Miyao, J. Tsujii. Probabilistic disambiguation models for wide-coverage HPSG parsing. *Proceedings ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* 2005;:83–90.
[24] K. Sagae. Dependency parsing and domain adaptation with LR models and parser ensembles. *In proceedings of the eleventh conference on computational natural language learning* 2007;:1044–1050.
[25] G. Minnen, J. Carroll, D. Pearce. Applied morphological processing of English. *Natural Language Engineering* 2001;7:207–223.
[26] S. Pyysalo, T. Salakoski, S. Aubin, A. Nazarenko. Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches. *BMC Bioinformatics* 2006;**7**:S2.
[27] L. Hunter, Z. Lu, J. Firby, W.A. Baumgartner, H.L. Johnson, P.V. Ogren, et al. OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics* 2008;**9**:78.
[28] Y. Tsuruoka. Bidirectional inference with the easiest-first strategy for tagging sequence data. *In proceedings of human language technology conference and conference on empirical methods in natural language processing* 2005;:467–474.